# Decision Trees

## Letter Recognition Using Holland-Style Adaptive Classifiers [1]

PETER W. FREY Department of Psychology, Northwestern University, Evanston, IL 60208 DAVID J. SLATE Pattern Recognition Group, Odesta Corporation, 1890 Maple Avenue, Evanston, IL 60201

Original paper

* Dataset features were scaled to 1-15.
* 4000 testing, 16000 training
* 5 fold
* Last training on whole set, one pass to further tune: prune tree rules that were below a level were removed.
* Each classifier is important
* Last test uses 4000 test items
* Each category has 150 rules
* Fixed size rule buffer of 3900 (max num splits?)
* Utility criterion set to 1/8000 gives best result
* exemplar-based rule creation produced good candidates at a much faster rate than the random or hybrid,
* Our results with the character recognition task indicated that reward sharing is much more effective than a winner-take-all.

## DECISION TREE LEARNING BASED FEATURE EVALUATION AND SELECTION FOR IMAGE CLASSIFICATION [2]

HAN LIU, MIHAELA COCEA, WEILI DING,

* 70:30 split
* 10 folds
* Rules:
  + Confidence is a measure of the weight (predictive accuracy)
  + J-measure is also one of the most popular way of ranking rules, cross entropy.

# Naïve Bayes Documents

## Multilabel naïve Bayes classification considering label dependence [3]

*Hae-Cheon Kim, Jin-Hyeong Park, Dae-Won Kim, Jaesung Lee∗*

**Good overview of different algorithms**

, the conventional multilabel naïve Bayes classifier treats each label independently and hence neglects the relations among labels, resulting in degenerated accuracy. For example, in the weather classification problem, the label raining is likely to be coupled with the label cloudy and unlikely to be coupled with the label sunny.

In this paper, we propose a new multilabel naïve Bayes algorithm that considers the dependence among labels for the classification process, named MLNB-LD

e label dependence can be broadly divided into three groups according to how many labels are considered concurrently.

1. The first group of classifiers treats each label independently by inferencing a mapping function for each label. For example, Zhang and Zhou.
2. In the second group, multilabel classifiers consider the label dependence between label pairs. For example, Huang et al.
3. In the third group, the classification process is designed to consider an arbitrary number of labels concurrently.

We presented a multilabel naïve Bayes classifier that considers the dependence among labels during classification. The proposed method utilizes the dependence between label pairs for determining the most probable label set for a given unseen instance. Our comprehensive experiments demonstrate that multilabel classification performance can be significantly improved by the proposed method. A comparison of the results obtained on 14 realworld datasets obtained from different domains shows the advantages of the proposed method compared with the four conventional multilabel classifiers in terms of three evaluation measures

## A versatile framework for labelling imagery with a large number of

## Classes[4]

*Shailesh Kumar', Melba Crawford2 and Joydeep Ghoshl; Department of Electrical and Computer Engineering University of Texas at Austin*

**Refers to letter dataset**

A Bayesian pairwise classifier architecture for a large number of classes was proposed and evaluated in this paper. This architecture is novel in three ways. First it provides a framework for pair-class feature selection that was not available in the conventional feature selection techniques. Second, the evaluation criteria uses log odds measure on class aposteriori probabilities to advantage. Third, a fast incremental algorithm is used for learning MOG’s in higher dimensional feature space using the MOG’s in the lower dimensional feature subspaces, based on knowledge reuse methods.

## A sequential feature extraction approach for naïve bayes classification of microarray data[5]

Liwei Fan a,\*, Kim-Leng Poh a , Peng Zhou

**Explains feature selection can improve performance for large datasets, and address attribute dependence: naïve bayes is implemented to treat attributes as independent.**

Nevertheless, there exist two major limitations that may severely affect the successful application of naïve Bayes classifier to microarray data analysis (Fan & Poh, 2008). The first is the conditional independence assumption embedded in the classifier itself, which is hardly satisfied by the microarray data. This limitation could be, at least theoretically, overcome by the class-conditional independent component analysis (CC-ICA) technique proposed in Vitria, Bressan, and Radeva (2007). Previous experimental results have shown that CC-ICA could effectively improve the performance of naïve Bayes classifier in some application domains.

Another limitation comes from the intrinsic characteristics of microarray dataset, which usually consists of thousands of genes with only tens of samples due to the expensive experiment. The extremely high dimensionality of microarray data may greatly increase the computational costs of naïve Bayes classifier. In addition, since the sample size is far less than the gene size, the use of CC-ICA can hardly enhance the independence among genes as well as improve the performance of naïve Bayes classifier

# Random Forest Documents, Decision Trees

## An Empirical Study of Learning from Imbalanced Data Using Random Forest[6]

Taghi M. Khoshgoftaar (taghi@cse.fau.edu) Moiz Golawala (moizgolawala@hotmail.com) Jason Van Hulse (jvanhulse@gmail.com) Department of Computer Science and Engineering Florida Atlantic University, Boca Raton, FL, USA

* Using 100 trees is best

## Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data [7]

Taghi M. Khoshgoftaar, Member, IEEE, Jason Van Hulse, Member, IEEE, and Amri Napolitano

The general conclusion obtained from our experiments is the clear preference of bagging over boosting when imbalanced data are noisy.

**Paper uses letter dataset**

## Random Decision Forests [8]

*Tin Kam Ho AT&T Bell Laboratories*

**Compares decision trees and random forests**

But trees derived with traditional methods often cannot be grown to arbitrary complexity for possible loss of generalization accuracy on unseen data.

The essence of the method is to build multiple trees in randomly selected subspaces of the feature space. Trees in different subspaces generalize their classification in complementary ways, and their combined classification can be monotonically improved.

In either method the stopping rule is until all the terminal nodes (leaves) contain points of a single class, or until it is impossible to split further. Since we do not want to lose any accuracy on classifying the training data, we do not consider methods to prune back the tree.

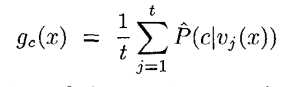
Our method to create multiple trees is to construct trees in randomly selected subspaces of the feature space. For a given feature space of m dimensions, there are 2" subspaces in which a decision tree can be constructed.

A decision tree is constructed in each selected subspace using the entire training set and the algorithms given in the previous section. Notice that each of these trees classifies the training data 100% correctly. Thus each tree generalizes its classification in a different way.

**Discriminant function**

For a point 2, let v3(x) be the terminal node that x is assigned to when it descends down tree T3 (j = 1,2, ..., t). Given this, let the posterior probability that x belongs to class c (c = 1,2, ..., n) be denoted by P(ClU3 (x)).

The discriminant function is defined as :



the decision rule is to assign x to class c for which gc(x) is the maximum.

**Single Trees**

First we show the results when single trees are constructed in the full feature space, as in the conventional practice. We tested both vectors fi and f2 and both of the tree growing methods.

Two heuristics used & compared: Central axis projection vs Perceptron training

**Random Forest**

F1 has 400 components. F2 has 852 components

The subspaces were restricted to 100 or 200 dimensions in the experiments, and the resultant differences in classification accuracy are clear from the figures.

## Shape Quantization and Recognition with Randomized Trees

*Yali Amit and Donald Geman* [9]

* Letter Recognition
* Looks Well Explained

## Random Forests, [6]

*LEO BREIMAN, Statistics Department, University of California, Berkeley*[10]

* Long and mathematical

## Letter Recognition Using Holland-Style Adaptive Classifiers,

*Frey & Slate*[1]

**Originator of dataset paper (decision tree).**

# Comparison of Algorithms

## Approximate Statistical Tests for Comparing Supervised Classication Learning Algorithms [11]

Thomas G. Dietterich tgd@cs.orst.edu Department of Computer Science Oregon State University

**Explains pitfalls of type 1 error (noise). 10-fold cross validation.**

This paper reviews five approximate statistical tests for determining whether one learning algorithm out-performs another on a particular learning task. These tests are compared experimentally to determine their probability of incorrectly detecting a difference when no difference exists (type I error).

* Two widely-used statistical tests are shown to have high probability of Type I error in certain situations and should never be used. These tests are (a) a test for the difference of two proportions and (b) a paired-differences t test based on taking several random train/test splits.
* A third test, a paired-differences t test based on 10-fold cross-validation, exhibits somewhat elevated probability of Type I error.
* A fourth test, McNemar's test, is shown to have low Type I error.
* The fifth test is a new test, 5x2cv, based on 5 iterations of 2-fold cross-validation. Experiments show that this test also has acceptable Type I error.

# Algorithm Evaluation

## Ensemble methods based on bias–variance analysis [12]

*Giorgio Valentini, Dipartimento di Informatica e Scienze dell’Informazione*

**Uses SVM’s so not sure**

bias–variance analysis is applied as a tool to explain the properties of ensembles of learners. A bias–variance analysis of ensembles based on resampling techniques is conducted, showing that, as expected, bagging is a variance reduction ensemble method,

## Mixtures of Conditional Maximum Entropy Models [13]

*Dmitry Pavlov (Yahoo inc.), Alexandrin Popescul ; University of Pennsylvania, David M. Pennock Overture Services Inc; Lyle H. Ungar, ; University of Pennsylvania*

**Not sure why I picked this paper, uses Maximum Entropy to analyze results of algorithms.**

## Normalization in sustainability assessment: Methods and implications [14]

# ML Papers

## An Introduction to Variable and Feature Selection[15]

*Isabelle Guyon, Andre Elisseeff*

Perfectly correlated variables are truly redundant in the sense that no additional information is gained by adding them.

Very high variable correlation (or anti-correlation) does not mean absence of variable complementarity.

References

[1] P. W. Frey and D. J. Slate, “Letter recognition using Holland-style adaptive classifiers,” *Mach. Learn.*, vol. 6, no. 2, pp. 161–182, Mar. 1991, doi: 10.1007/BF00114162.

[2] H. Liu, M. Cocea, and W. Ding, “DECISION TREE LEARNING BASED FEATURE EVALUATION AND SELECTION FOR IMAGE CLASSIFICATION,” p. 6.

[3] H.-C. Kim, J.-H. Park, D.-W. Kim, and J. Lee, “Multilabel naïve Bayes classification considering label dependence,” *Pattern Recognit. Lett.*, vol. 136, pp. 279–285, Aug. 2020, doi: 10.1016/j.patrec.2020.06.021.

[4] S. Kumar, M. Crawford, and J. Ghosh, “A versatile framework for labelling imagery with a large number of classes,” in *IJCNN’99. International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339)*, Washington, DC, USA, 1999, vol. 4, pp. 2829–2833, doi: 10.1109/IJCNN.1999.833531.

[5] L. Fan, K.-L. Poh, and P. Zhou, “A sequential feature extraction approach for naïve bayes classification of microarray data,” *Expert Syst. Appl.*, vol. 36, no. 6, pp. 9919–9923, Aug. 2009, doi: 10.1016/j.eswa.2009.01.075.

[6] T. M. Khoshgoftaar, M. Golawala, and J. V. Hulse, “An Empirical Study of Learning from Imbalanced Data Using Random Forest,” p. 8.

[7] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, “Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data,” *IEEE Trans. Syst. Man Cybern. - Part Syst. Hum.*, vol. 41, no. 3, pp. 552–568, May 2011, doi: 10.1109/TSMCA.2010.2084081.

[8] Tin Kam Ho, “Random decision forests,” in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, Que., Canada, 1995, vol. 1, pp. 278–282, doi: 10.1109/ICDAR.1995.598994.

[9] Y. Amit and D. Geman, “Shape Quantization and Recognition with Randomized Trees,” *Neural Comput.*, vol. 9, no. 7, pp. 1545–1588, Oct. 1997, doi: 10.1162/neco.1997.9.7.1545.

[10] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[11] T. G. Dietterich, “Approximate Statistical Tests for Comparing Supervised Classi cation Learning Algorithms,” p. 24.

[12] G. Valentini, “Ensemble methods based on bias–variance analysis,” p. 175.

[13] D. Pavlov, A. Popescul, D. M. Pennock, and L. H. Ungar, “Mixtures of Conditional Maximum Entropy Models,” p. 8.

[14] N. L. Pollesch and V. H. Dale, “Normalization in sustainability assessment: Methods and implications,” *Ecol. Econ.*, vol. 130, pp. 195–208, Oct. 2016, doi: 10.1016/j.ecolecon.2016.06.018.

[15] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *J. Mach. Learn. Res.*, vol. 3, p. 26, 2003.